Overview of the Third NTCIR Workshop

Noriko Kando
National Institute of Informatics
Noriko.Kando@nii.ac.jp

Abstract

This paper introduces the third NTCIR Workshop, which is the latest in a series of evaluation workshops designed to enhance research in information access technologies, including information retrieval, automatic text summarization, question answering, etc., by providing large-scale test collections and a forum for researchers. In the third Workshop, document collections were diversified in the aspects of length, genres, and languages. The focus of evaluation was also diversified from document-level retrieval to processing on units smaller than document and technologies supporting users to utilize information in the documents. The purpose of this paper is to serve as an introduction to the research described in detail in the rest of this volume.

Keywords: *evaluation, information access, information retrieval, text summarization, question answering, test collections, cross-lingual information retrieval, patent retrieval, Web retrieval.*

1 Introduction

The NTCIR Workshop [1] is a series of evaluation workshops designed to enhance research in information access (IA) technologies.

The aims of the NTCIR Workshop are:

- to encourage research in information access technologies by developing and providing the infrastructure for large-scale evaluation of information access technologies,
- to provide a forum for research groups interested in cross-system comparison and in exchanging research ideas, and
- to investigate (a) methodologies and metrics for evaluation of information access technologies, and (b) methods for constructing large-scale reusable test collections.

The primary component of the evaluation infrastructure is large-scale test collections reusable for experiments.

The term *information access* (IA) refers the whole process that users obtained relevant information, from document collections, to solve their problems. Traditionally document retrieval has been the core technology to support such process, and then the scope of IA technologies is being evolved and diversified to include technologies supporting users to utilize information from documents, such as information retrieval (IR), text summarization, question answering (QA), text mining, etc.

In the NTCIR, attention has been given to, but not limited to, Japanese and East Asian languages, but NTCIR has attracted international participation.

The third NTCIR Workshop selected five areas of research as "tasks":

1. Cross-Language Information Retrieval (CLIR),
2. Patent Retrieval (PATENT),
3. Question Answering Challenge (QAC),
4. Text Summarization Challenge (TSC), and
5. Web Retrieval (WEB).

This was the first workshop to include PATENT, QAC and WEB tasks. TSC has a new subtask of multi-document summarization, and CLIR includes a new subtask of multi-lingual CLIR. In the NTCIR, a new challenging task was termed a "challenge". However, we found that all five tasks contained new and challenging issues regarding technologies as well as their evaluation, and therefore each task was a "challenge" for both the participants and the task organizers.

The Section 2 describes the settings of the third NTCIR Workshop. Section 3 briefly introduces each task. More detailed descriptions can be found in the task overview papers [2-6]. The purpose of this paper is to serve as an introduction for the research described in detail in the rest of the working notes. The final section is a summary in which some thoughts on future directions are presented.

2 Organization and Participation

2.1 Organization

The third NTCIR Workshop was co-sponsored by

Table 1: Active Participating Groups of the Third NTCIR Workshop

Chungnam National University (Korea) & ETRI+ (Korea)
Carnegie Mellon University (USA)
Communication Research Laboratory+ (3 groups) (Japan)
CRL+ (Japan) & New York University (USA)
Fu Jen Catholic University (Taiwan ROC)
Hitachi* (Japan)
Hong Kong Polytechnic University (Hong Kong)
Hummingbird* (Canada)
Institute of Software, Chinese Academy of Sciences+ (China, PRC)
Johns Hopkins University (USA)
Keio University (2 groups) (Japan)
Kent Ridge Digital Labs+ (Singapore)
Kochi University of Technology (Japan)
Korea University (Korea)
Matsushita Electric Industrial* (Japan)
Microsoft Research Asia* (China PRC)
Mie University (Japan)
Nara Advanced Institute of Science and Technology (Japan)
NAIST & CRL+ (Japan)
National Taiwan University (Taiwan ROC)
NEC Kansai* (Japan)
NEC MRL* (Japan)
NTT Data Technology* (Japan)
NTT-CS* (Japan)
NTT-CS* (Japan) & NAIST (Japan)

NTT DATA* (Japan)
New York University (USA) & CRL+ (Japan)
Oki Electric* (Japan)
Osaka Kyoiku Univeristy (3 groups) (Japan)
POSTECH (2 groups) (Korea)
Queen College City University of New York (USA)
RICOH* (Japan)
Ritsumeikan University (2 groups) (Japan)
SICS+ (Sweden)
Surugadai University (Japan)
Thomson Legal and Regulatory* (USA)
Tianjin University (China PRC)
Tokyo Institute of Technology (Japan)
Tokai University & Beijin Japan Center (China PRC)
Toshiba* (Japan)
Toyohashi University of Technology (4 groups) (Japan)
ULIS & AIST+ (2 groups) (Japan)
University Aizu (2 groups)
University of California Berkeley (2 groups) (USA)
University of Tokyo (2 groups) (Japan)
University of Lib and Information Science (2 groups) (Japan)
University of Tokyo (Japan) & RICOH* (Japan)
Waterloo University (Canada)
Yokohama National University (2 groups) (Japan)

65 groups from 9 countries, *: company, +: national or independent research institute, without-symbol: university

MEXT [1] *Grant-in-Aid for Scientific Research on Informatics (#13224087)*[2] and *the Research Center for Information Resources (RCIR) at the National Institute of Informatics (NII).*

Each task has been organized by the task organizers listed on the "Organization" page in the proceedings. They are researchers of the area and belong to the *NTCIR Research and Organizing Committee*. The committee had monthly meetings at NII and discussed the plan and problems related to the evaluation and task organization, and frequently discussed through emails. The CLIR task organizers consisted of members from Taiwan, Korea and Japan, and met four times at NII for discussion on task planning and schedule, dry-run planning, formal run topic selection and evaluation.

For the third NTCIR Workshop, the process started from the document data distribution in September 2001 and the workshop meeting was held on 8-10 October, 2002.

2.2 Participants

Table 1 is a list of the active participating research groups in the third NTCIR Workshop. Sixty-five groups from nine different countries and areas submitted task results. Among these, 14 groups are from companies, seven are from national or independent research institutes, and 44 are from universities. Collaborating research groups from different organizations are listed under the first organization.

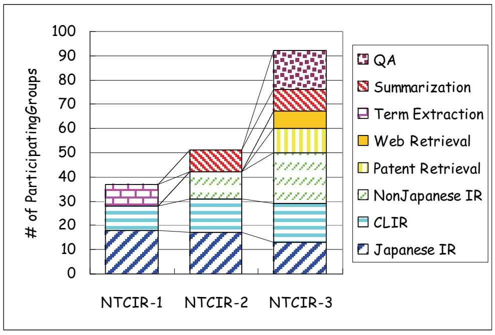

Figure 1: Number of Participating Groups by Tasks

[1] Japan's Ministry of Education, Culture, Sports, Science and Technology

[2] "Digital Contents Research Group A02 of Japanese MEXT Grant-in-Aid for Scientific Research on Informatics (#13224087), (Principal Investigator : Jun Adachi)

Table 2: Test Collections Constructed through NTCIR

collection	task	documents			topic		relevance judgment
		genre	size	lang	lang	#	
NTCIR-1	IR	sci. abstract	577MB	JE	J	83	3 grades
CIRB010	IR	newspaper 98-9	210MB	C	CE	50	4 grades
NTCIR-2	IR	sci. abstract	800MB	JE	JE	49	4 grades
NTCIR-2 SUMM	Summ	newspaer94,95,98	180 doc	J	J	-	-
NTCIR-2TAO	Summ	newspaer98	1000 doc	J	J	-	-
KEIB010	IR	newpaper94	74MB	K	CKJE	30	4 grades
CIRB011+020, NTCIR-3CLIR	IR	newspaper 98-9	870MB	CJE		50	4 grades
NTCIR-3PAT	IR	patent full'98-9	17GB	J	CCKJE	31	3 grades
		+abstract'95-9	4GB	JE			
NTCIR-3 QA	QA	newspaper 98-9	282MB	J	J	240+60+ about 900	2 grades
NTCIR-3 SUMM	Summ	newspaper 98-9	30 docs + 30 sets of docs	J	J	-	-
NTCIR-3Web	IR	HTML	100GB	J(E)	J	47	5 grades

J:Japanese, E:English, C:Chinese, K:Korean

As shown in Figure 1, increasing the variety of tasks and languages attracted many newcomers from various research communities. Some of these are experienced TREC [7] participants in cross-lingual information retrieval and question answering. The PATENT and WEB tasks use larger document collections with different structures, and PATENT task attracted "veteran" NTCIR participants and participants from company research laboratories. The number of collaborating groups from different organizations increased.

The collaboration of research groups from different technology areas had interesting effects on the task results and contributed to an enhanced variety of approaches and strategies. For example, University of Tokyo and Ricoh Joint team, which participated in WEB task, is a collaboration of an experienced text retrieval research group both in TREC and NTCIR and an experienced research group on link analysis for the Web. This collaboration resulted in a retrieval strategy combining content-based text retrieval and link-based retrieval, and it worked well. We hope that such fruitful collaborations across the research groups will increased through NTCIR.

2.3 Test Collections

Table 2 shows the test collections constructed through the series of NTCIR workshops [8-10] and Table 3 shows the test collections used in each task at the third NTCIR workshop. All the documents data were provided to the participants of the task from the NII free of charge after exchanging the user agreements. After the workshop, NTCIR-3 PATENT, NW100G-01, and NW10G-01 are available from the RCIR in the NII for research purpose use to any researchers. Topics and relevance judgments, questions and answers, and summaries produced by human professionals as referenced data are also available from the RCIR in the NII for research purpose use. The news articles of Mainichi Newspapers are available from *Nichigai Associates Co*. for research purpose use with charge but document data in CIRB011, CIRB020 and KEIB010 are currently only available for the Workshop participants.

2.3.1 Documents

In the third NTCIR Workshop, we used three different document genres as document collections;

1. news articles (CLIR, QAC, and TSC),
2. patents (PATENT), and
3. Web documents (WEB).

CLIR task used news articles published in Taiwan, Korea and Japan in own languages and English articles published in Taiwan and Japan. The Japanese news articles used in CLIR task, QAC, and TSC are the same document files of the *Mainichi Newspaper 1998-1999.*

The Patent collection, *NTC-3 PATENT*, consists of the full text Japanese patents published in 1998-1999 and Japanese and English exactly translated abstracts published in 1995-1999. All the documents were provided by the PATOLIS Corporation to the NTCIR project for research purpose use in NTCIR. For the cross-lingual information retrieval using this collection, when the documents published in 1998-1999 are retrieved, the abstracts publish in 1995-1997 are usable to extract translation knowledge. PATENT task investigated "Cross-genre" retrieval from new articles to patent, and each of the topics (search requests) used in the task contains a news article selected from the *Mainichi Newspaper 1998-1999.*

WEB documents contain text, meta data, html tags and links. They were crawled mainly from ".jp" (Japan) domain in 2001 and those written in Japanese or English were judged relevance. The contents of the NW100G-01 and NW100G-01 can be accessed on the servers in the "Open Lab" at the NII. Each participant provided a workstation and disk space in the Open Lab, then access to these facilities physically (i.e. come to the Open Lab) or via network. The document data can not be downloaded or copied to any place outside the Lab but the index or processed data from the documents can be bring out from the Open Lab to own working place if the original documents can not be reproduced from them.

Table 3: Tasks and Test Collection Used in the Third NTCIR Workshop

	period	tasks	subtasks	test collections
3	Sept. 2001- Oct. 2002	CLIR	single lang IR: C-C,K-K,J-J	NTCIR-3CLIR, CIRB010, CIRB020, KEIB010
			bilingual CLIR: x-J,x-C, x-K	
			mulilingual CLIR: x-CJE	
		Patent Retrieval	cross genre	NTCIR-3Patent
			CLIR: x-J, x-JabstEabst	
			optional task: alignments, readability	
		Question Answering	task1- 5 candidate answers	NTCIR-3QA
			task2-one set of all the answer	
			task3-series of questions	
		Text Summarization	task A: single text	NTCIR-3Summ
			task B: multiple texts	
		Web Retrieval	survey retrieval	NW100GB, NW10GB
			target retrieval	
			optional task: search result classification, speech-driven retrieval	

"n-m" for CLIR: n=query language, m=document language(s), J:Japanese, E:English, C:Chinese, K:Korean, x:any of CJKE

Figure 2 shows a sample document record in the CIRB011. Document records in all the test collections listed above are plain texts with SGML-like tags. In *NW100G-01* and *NW10G-01*, a text region between <HTML> and </HTML> contains a html document crawled and it keeps original html tags.

Encodings are Big5 for (Traditional) Chinese documents in *CIRB011* and *CIRB020*, EUC for Korean documents in *KEIB010* and EUC for Japanese documents in *Mainichi*, *kkh*, and *jsh.* For *NW100G-01 and NW10G-01,* three types of document files were available for the participants, i.e., (1) original encodings as crawled (Japanese documents may be encoded by EUC, JIS, Shift JIS or Unicode), (2) All the Japanese documents were converted into EUC, and (3) the documents in which html tags were discarded from (2).

In a series of NTCIR Workshops, each workshop chose a particular document genre or genres as the document collections used for the experiments.

```
<DOC>
<DOCNO>ctg_xxx_19990110_0001</DOCNO>
<LANG>EN</LANG>
<HEADLINE> Asia Urged to Move Faster in Shoring Up
Shaky Banks </HEADLINE>
<DATE>1999-01-10</DATE>
<TEXT>
<P>HONG KONG, Jan 10 (AFP) - Bank for International
Settlements (BIS) general manager Andrew Crockett has
urged Asian economies to move faster in reforming their
shaky banking sectors, reports said Sunday. Speaking
ahead of Monday's meeting at the BIS office here of
international central bankers including US Federal
Reserve chairman Alan Greenspan, Crockett said he
was encouraged by regional banking reforms but "there
is still some way to go." Asian banks shake off their
burden of bad debt if they were to be able to finance
recovery in the crisis-hit region, he said according to the
Sunday Morning Post. Crockett added that more stable
currency exchange rates and lower interest rates had
paved the way for recovery. "Therefore I believe in the
financial area, the crisis has in a sense been contained
and that now it is possible to look forward to real
economic recovery," he was quoted as saying by the
Sunday Hong Kong Standard.</P>
<P>"It would not surprise me, given the interest I know
certain governors have, if the subject of hedge funds was
discussed during the meeting," Crockett said. </P>
<P>He reiterated comments by BIS officials here that the
central bankers would stay tight-lipped about their
meeting, the first to be held at the Hong Kong office of the
Swiss-based institution since it opened last July. </P>
</TEXT>
</DOC>
```

Figure 2: Sample Document (*CIRB011*)

It is because we would like (1) to increase the variety of document genres usable for experiments, (2) to investigate problems and applications appropriate to each document genre, and (3) to investigate explicitly the technologies to overcome the heterogeneity of the document genre(s) that found in the operational setting. And then we have tried to design the experiments and evaluation based on the users' information tasks using the documents of the genre. The relevance judgements are all done as graded-judgments. CLIR and WEB used 4 grades and PATENT used 3 grades. Judging in 3 grades is natural for the real users of the patent retrieval systems and appropriate in their operational and ordinary usage.

This was the first NTCIR workshop that the scientific document collections from the NII's *NACSIS-IR* service are not used in any of tasks. The decision was made by conjunction of several reasons and conditions.

2.3.2 Topics and Relevance Judgments for IR

Topics

A sample topic record used in the CLIR at the NTCIR Workshop 3 is shown in Fig. 2. Topics are defined as statements of "user's requests" rather than "queries", which are the strings actually submitted to the system because they are usable for both manual and automatic query construction.

```
<TOPIC>
<NUM>013</NUM>
<SLANG>CH</SLANG>
<TLANG>EN</TLANG>
<TITLE>NBA labor dispute</TITLE>
<DESC>
To retrieve the labor dispute between the two parties
of the US National Basketball Association at the end
of 1998 and the agreement that they reached.
</DESC>
<NARR>
The content of the related documents should include
the causes of the NBA labor dispute, the relations
between the players and the management, main
controversial issues of both sides, compromises
after negotiation and content of the new agreement,
etc. The document will be regarded as irrelevant if it
only touched upon the influences of closing the court
on each game of the season.
</NARR>
<CONC>
NBA (National Basketball Association), union, team,
league, labor dispute, league and union, negotiation,
to sign an agreement, salary, lockout, Stern, Bird
Regulation.
</CONC>
</TOPIC>
```

Figure3: A Sample Topic (NTCIR-3 CLIR)

An NTCIR topic used for IR related tasks contains four basic sets or fields, *i.e.*, the title of the topic (<TITLE>), a brief description (<DESC>), a detailed narrative (<NARR>), and a list of concepts (<CONC>). <TITLE> can be used as a very short query resembling one often submitted to search engines, although it may not cover all the major concepts of the search request. <DESC> basically contains all the major concepts of the search request. <NARR> may contain term definitions, background knowledge, the purpose of the search, criteria for relevance judgments, etc. It is known that the runs using it tended to attain very high search effectiveness. Attention must be paid to comparing runs with and without <CONC>.

- *Query types*

Any fields of the NTCIR topics can be used in the retrieval. However, each task set a "*mandatory run*" using only a particular field(s) of the topics in the retrieval and every participant is requested to submit at least one such set of retrieval results. The purpose of this is to enhance the cross-system comparison on the common setting. "*D-run*", a run using <DESC> only was set as mandatory in the previous NTCIR workshops. This may vary with the purpose of evaluation and task design.

Topics can extend this basic structure and be designed in accordance with the design and purpose of the task. For example, topics used in PATENT contain many additional fields such as <ARTICLE> and <SUPPLEMENT> for cross-DB retrieval. <TITLE> in the NTCIR-3 WEB topics is specially designed to mirror the ways in which ordinary Web search-engine users submit queries.

- *Query methods*

Participants can use any method to create queries from the topic statements. In NTCIR, "automatic" is a query construction without any human intervention, and "manual" is any method other than automatic. Therefore, "manual" runs may include a wide variety of levels of human intervention in query construction.

Any experimental results using a test collection must be reported together with the conditions of the query fields and query methods used. This is because retrieval effectiveness can vary depending on these conditions.

Relevance judgments (Right answers)

The relevance judgment is a list of documents in a particular document collection that are relevant to a particular topic, and they are the "right answers" for retrieval tasks. With such "right answers", a document collection and set of topics becomes a "test collection" for retrieval tasks.

The criterion for the success of a search is "*relevance*"—the judgments of a human assessor (who acts as a user of the retrieval system) whether the retrieved document contains relevant information to his/her information needs described as a format of a "topic". "*Relevance*" is one of the most central concepts in information retrieval, and various aspects of it have been discussed by various researchers in [19-23].

Relevance is completely different from "matching" the query term with the terms in the documents, and deals with the "information" described in the documents. Therefore, a document containing a topic term can be judged as irrelevant and a document without the topic terms can be judged relevant if it contains relevant information. This makes IR both difficult and interesting.

In the NTCIR, relevance judgments were conducted using multi-grades because we thought it more natural, although the evaluation increased in complexity [24-26].

An assessment system was developed and used in CLIR, PATENT, and WEB[3]. The system is relatively flexible and can be set to link to other documents. This function is critical for judgments for the WEB and PATENT tasks. Relevance judgment files not only contained the relevance of each document in the pool but also contained extracted phrases or passages showing the reason the analyst assessed the document as "relevant". These statements were used to confirm the judgments and also retained for future use in experiments related to extracting answer passages.

3.2.2 Question Answering

Question answering (QA) is a form of extension of IR and information extraction (IE), combining both technologies. The QA track at TREC 8 is the first venture to address the technology and construction of a test collection for experimentation. QA is a technology extracting "an answer" to a particular question, rather than documents containing "answers", from knowledge sources such as document collections etc. For a collection to be usable for QA evaluation, it must have at least three components similar to the IR test collections:

(1) a set of "knowledge sources"—document collections etc.,
(2) a set of questions, and
(3) lists of the answer(s) to each question.

In order to evaluate QA technologies, there are several technical aspects concerning the extraction of answer expressions from knowledge sources, including (i) Question type, (ii) number of answer expressions in knowledge sources, (iii) types of answers extracted, etc.

In QAC at the NTCIR-3, three types of sub-tasks were selected and in each of them a set of 5W1H-type questions were asked on various subject domains and various categories of named entities. QAC at NTCIR-3 test set allowed questions with more than one answer or no answer and uses *exact answers*—nouns or noun phrases rather than short passages including answers.

Sample questions are shown in Fig. 3. For

[3] We thank Ellen Voorhees who kindly provided the TREC's assessment system source code. Although because of the differences of the environment, we have developed our assessment system from the scratch, we have learned the design concepts and requirements from the TREC's system and their experience.

example, the first question in it requires a set of multiple answers, *i.e.* "DDI, IDO, and KDD".

QAC1-1001-01：″２０００年１０月１日に合併することが決まった通信三社はどこですか。″
QAC1-1002-01：″広辞苑第五版はいつ発売されましたか。″
QAC1-3003-01：″ＮＨＫ連続テレビ小説の平均視聴率は最高どのくらいですか。″
QAC1-3003-02：″それをとったドラマのタイトルは何ですか。″

(English translation).
QAC1-1001-01： ″Which three telecommunications companies decided to merge on October 1, 2000?″
QAC1-1002-01： ″When was the fifth edition of the *Kojien* Japanese dictionary published?″
QAC1-3003-01：″What is the highest record average audience rating for NHK's morning TV Novel Series?″
QAC1-3003-02：″What is the title of the drama that set the highest record?″

Figure3: Sample Questions (QAC1)

There has been a trend towards evaluation using more realistic types of question and answer, and TREC QA tracks [27-29] have worked in that direction. NTCIR QAC added to the variety of QA test collections available.

The "reusability" of the QA test collections is still challenging. For example, when a QA system produces answers that are passages that include the answers, rather than exact answers, automatic assessment of the results can cause problems [29]. Automatic assessment of the results can be performed for QA systems producing exact answers. However, QA technologies are easily tuned to the particular collection used in the experiments unless there is a substantial number of topics. Careful consideration of this aspect must take place both for the experiments and for the comparison of the results.

3.2.3 Text Summarization

Research on automatic text summarization has been done since the 1950s, but the issue of how to evaluate it has not been discussed until recently. Through SUMMAC and DUC, discussion and investigation of the evaluation methodologies of text summarization was enhanced greatly.

For example, a data set used for extrinsic evaluation, such as an IR task-based one, can be similar to the IR test collection. Evaluation of the system-produced summaries is done by the ratio of the consistency of the relevance judgments between judges done on the original texts and those on the summaries.

In intrinsic evaluation, collected human-created summaries are often usable in various ways, including as key reference data in evaluation, etc. In this case, the test collection for text summarization consists of (1) documents, (2) instruction of summary production, including the specified length of the target summary, summarization types such as “extract” or “abstract”, (3) a set of hand-created summaries used for references. It has been recommended that more than one summary for each source document or document set prepared by different analysts be used. TSC2 in NTCIR-3 used three different summarization professionals to produce summaries for each document or document set to be summarized. We also newly added summaries that were produced by two different analysts to the collection of NTCIR-2 Summ, which was used at TSC in NTCIR Workshop 2.

For multi-document summarization at TSC2 in NTCIR-3, a "topic" for each set of documents was provided. It indicated the focus of the summary produced from the set of documents and helped the systems or users to select or focus which part of the documents should be summarized. It can assume a topic of a search request and a task summarizing the contents of the set of retrieved documents relevant to the topic.

Such collections of hand-created summaries are usable for summarization research in various ways. It is difficult to make a reusable collection for intrinsic summarization evaluation, and there are still many opportunities to investigate and discuss the evaluation design for summarization.

3.2.4 Linguistic analysis

NTCIR-1 contains a "*Tagged Corpus*". This contains detailed hand-tagged part-of-speech (POS) tags for 2,000 Japanese documents selected from NTCIR-1. Spelling errors are manually collected. Because of the absence of explicit boundaries between words in Japanese sentences, we set three levels of lexical boundaries (i.e., word boundaries, and strong and weak morpheme boundaries). This was originally constructed for the Term Extraction and Role Analysis Task at the first NTCIR Workshop.

In NTCIR-2, another type of segmented data for the complete Japanese document collection and topics was provided in Information Retrieval task. They are segmented into three levels of lexical boundaries using a commercially available morphological analyzer called HAPPINESS. An analysis of the effect of segmentation is reported in [30]. No particular linguistic analysis data was newly developed for NTCIR-3.

3.2.5 Robustness of the system evaluation using the test collections

The test collections NTCIR-1 and -2 have been tested for the following characteristics, to enable their use as a reliable tool for IR system testing:

- exhaustiveness of the document pool
- consistency between assessors and its effect on system evaluation
- topic difficulty

The results have been reported and published on various occasions [30-37]. In terms of exhaustiveness, pooling the top 100 documents from each run worked well for topics with fewer than 100 relevant documents. For topics with more than 100 relevant documents, although the top 100 pooling covered only 51.9% of the total relevant documents, coverage was higher than 90% if combined with additional interactive searches. Therefore, we conducted additional interactive searches for the topics with more than 50 relevant documents in the first workshop, and those with more than 100 relevant documents in the second workshop.

When the pool size was larger than 2,500 for a specific topic, the number of documents collected from each submitted run was reduced to 80 or 90. This was done to keep the pool size practical and manageable for assessors to keep consistency in the pool. Although the numbers of documents collected in the pool were different for each topic, the number of documents collected from each run was exactly the same for a specific topic.

A strong correlation was found to exist between the system rankings produced using different relevance judgments and different pooling methods, regardless of the inconsistency of the relevance assessments among analysts and regardless of the different pooling methods used [31-33, 35]. This served as an additional support to the analysis reported by Voorhees [38].

3.2.6 Continuous effort for enhancement

TREC maintained the long tradition of text retrieval test collections starting with Cranfield projects in the 1960's [7, 14-17] and spun out CLEF [18] , and has enhanced and evolved in various ways, making them more realistic by responding to the needs of the social and technological environments of current society, which are continuously improving and changing.

For example:

(a) using written statements of user information requests as topics created by users, and judging the relevance based on the topic statements rather than queries, which are the strings input to the systems by users

(b) scaling up the document collection size to be comparable with the operational setting

(c) enhancing the scope beyond "English text" in the types of languages and media, such as OCRed texts, spoken documents, video

(d) enhancing the scope of the technologies beyond "document retrieval", for example question answering

(e) incorporating Web documents, the most common document type today and quite unique

Technologies keep improving. The evaluation methodologies and metrics must continuously improve and change in response to social and technological needs.

3.3 Evaluation

Task results on a test collection can be evaluated in various ways.

Results of the retrieval tasks at NTCIR are evaluated using the *trec-eval* program, which was written by Chris Buckley [39]. This can provide 85 different scores for a run, and reported scores for each topic, as well as a score averaged over all topics. Among these, the Recall-Precision curve graph and Mean Average Precision over non-interpolated all relevant documents are the most common measures used to report the NTCIR evaluation results. The former shows the balance of the runs and the latter is a very stable measure when an adequate number of topics is used.

Discounted Cumulated Gain [24] and Mean Reciprocal Rank [28] were used in the WEB and Task 1 of QAC1. F-measure is also used in Task 2 and 3 of QAC1. TSC2 used subjective, intrinsic evaluation: (1) evaluation using "ranking" by human assessors, who are shown several summaries and assign the rank considering content and readability, and (2) evaluation by revision, *i.e.*, how many edits (deletions, insertions and replacements) were performed by human assessors on the system-produced summary.

In addition, new measures, including novelty-based evaluation for PATENT, and weighted average precision for multi-grade relevance judgments, were proposed. However, they were not used in the formal evaluation of the tasks at the NTCIR Workshop 3.

4 The Third NTCIR Workshop

This section provides brief introductions to each task design.

4.1Cross-Language Retrieval Task (CLIR)[2]

The purpose of the task is to evaluate a much more complex CLIR evaluation task, which is closer to realistic applications in the IR environment and is a real challenge to IR researchers. It was organized by the nine researchers from Japan, Korea, and Taiwan They met three times in Japan to discuss the details of the CLIR Task, to determine the schedule, and to arrange the agenda.

Topic creation and relevance judgments on each language document set were done by each organizers in of the language based on the common Topic Creation Manual and Relevance Judgment Instruction. Task description announcement, evaluation and report writing were done by Kuang-hua Chen, and pooling was done by Kazuko Kuriyama. Kazuaki Kishida did additional analysis and coordination.

4.1.1 Subtasks

The CLIR task set three subtasks;

1. *Multilingual CLIR (MLIR)*: Search the document collection of more than one language (Xtopic98>CEJ).
2. *Bilingual CLIR (BLIR)*: CLIR between any two different languages, except the run on English documents (Xtopic98>C, Xtopic94>K, Xtopic98>J).
3. *Single Language IR (SLIR)*: Monolingual IR (Ctopic98>C, Ktopic94>K, Jtopic98>J).

4.1.2 Test Collections

There are two sets according to the publication year:

- 1998–99 Set: Chinese, Japanese, and English documents, and 50 topics (Topic98).
- 1994 Set: Korean documents and 30 topics (Topic94).

Both topic sets contain four languages: Chinese, Korean, English and Japanese. A sample document and topic are shown in Figs. 1 and 2.

Relevance judgments were presented in four grades: Highly relevant (S), Relevant (A), Partially relevant (B), and Irrelevant (C).

On the "layers of CLIR technologies" [40], the CLIR of newspaper articles related to the "pragmatic layer (social, cultural convention, etc)", and cultural/social differences are the issues in both topic creation and retrieval. The remaining local and international topics and topics including proper names were considered.

NTCIR-3 Formal Test Collection for CLIR was selected based on the "*3-in-S+A*" criterion: a qualified topic must have at least three relevant documents with 'S' or 'A' score on each document collection as follows:

(1) *NTCIR-3 Formal Chinese Test Collection:* 381,681 documents and 42 topics from Topic98.

(2) *NTCIR-3 Formal Japanese Test Collection:* 220,078 documents and 42 topics from Topic98

(3) *NTCIR-3 Formal English Test Collection:* 22,927 documents and 32 topics from Topic98.

(4) *NTCIR-3 Formal CJ Test Collection:* 601,759 documents and 50 topics from Topic98.

(5) *NTCIR-3 Formal CE Test Collection:* 404,608 documents and 46 topics from Topic98.

(6) *NTCIR-3 Formal JE Test Collection:* 243,005 documents and 45 topics from Topic98.

(7) *NTCIR 3 Formal CJE Test Collection:* 624,686 documents and 50 topics from Topic98.

(8) *NTCIR-3 Formal Korean Test Collection:* 66,146 documents and 30 topics from Topic94.

Twenty groups from eight countries submitted results; 110 runs for SLIR, 50 for BLIR and 29 for MLIR. A good number of groups participated in every language SLIR and investigated in detail appropriate retrieval strategies for each language, including comparison of the effectiveness of segmentation strategies on each language, comparison of the retrieval models of Okapi and Pircs, logistic regression and vector space, etc. This direction of investigation provides a foundation for further research on MLIR in future workshops.

For BLIR, the retrieval effectiveness of C-J is rather low compared to E-C and E-J. This was partly because of less experience on CLIR among Asian languages compared to CLIR between English and the investigator's own language, and partly because of the lower availability of the resources for translation.

It was the first year of distributed topic creation. Both topic translation and relevance judgments are difficult for topics including local news or local proper names. To complement this problem, the topic—and especially <DESC>, which was used as the mandatory run—became longer compared to the previous NTCIR topics. Several strategies for disambiguation were used, but they did not test effectiveness on longer topics adequately.

For the next NTCIR Workshop, Korean newspaper articles published in 1998–99 in both English and the Korean language will be added, then MLIR of Chinese, Korean, Japanese and English that is published in Asia and will be feasible. In response to the above issues, the following are possible future directions of enhancement:

(1) Encourage participation in MLIR
(2) Pivot language subtask in MLIR; MLIR using English as a bridging language
(3) Prepare shorter <DESC> or set <TITLE> as mandatory
(4) Challenging tasks such as CLQA.

For (2), richer resources are available for CLIR between English and each language, and it can be considered as a realistic approach when the number of languages increases.

4.2 Patent Retrieval Task (PATENT) [3]

The purposes of PATENT task are (1) to enhance research on patent information processing by providing test collections for patents—from patent retrieval to patent mining, from technical survey to finding conflicting applications, monolingual and cross-lingual—and (2) to provide a test bed of information access including real task, real user, and real information needs for a variety of tasks (IR, CLIR, SDI, summarization, mining) [3]. The possibility and need for the patent retrieval task has been proposed since the first NTCIR Workshop because of characteristics of patent applications and the acute social need for advanced functionality and CLIR. This unique task was made feasible through the cooperation of various sectors. Patolis, Co., Information Retrieval Committee at the Japan Intellectual Property Association (JIPA), and IR researchers with experience in patent retrieval were task organizers.

4.2.1 Subtasks

1. Main Task:
   - Cross-DB (genre) Retrieval: retrieve patents in response to newspaper articles associated with technology and commercial products.
   - CLIR: search Japanese patents by any topic field with English, Chinese, or Korean topics.
2. Optional task:
   Proposal-based free-styled task.

4.2.2 Test Collections

(1) Japanese patents: 1998–1999 (17GB, 690K docs)
(2) JAPIO patent abstracts: 1995–1999 (1706K docs)
(3) Patent Abstracts of Japan (English translations of 2): 1995–1999 (ca. 1706K docs)
(4) 31 topics including newspaper articles, in Japanese, and translation into Chinese (traditional, simplified), Korean, English, and
(5) relevance judgments in three grades.

Organization of the document collection is shown in Fig. 1, and a sample topic in Fig. 2, in [3]. Patent documents have various unique characteristics as an application of text retrieval, including (1) structure, (2) complicated and vague sentences, (3) document length, (4) collection size, and (5) various alignments including English–Japanese paired abstracts as well as among the sections in a patent etc.

The organizers and patent professionals at JIPA discussed the task design so that:

- it is realistic as an operational patent application,
- the task can initiate a new direction of IR that can be applicable to other document genres, and
- it is feasible for the document collection provided.

They chose the context of the experimental design of "search for technological trend survey". Regarding "Cross-DB Retrieval", we assumed that someone sent a newspaper article clip to a patent intermediary and asked for the related patents to be retrieved. <ARTICLE> in the topic is a clip and <SUPPLEMENT> is a memo indicating the focus of the search in the article. Searches using ordinary topic fields such as <DESC>, <NARR>, etc. were accepted as non-mandatory runs. Association retrieval among patents can be done using the patent indicated in <PI>.

Topic creation and relevance judgments were conducted by professionals at Japan Intellectual Property Association (JIPA). Eight groups from three different countries submitted results for the main tasks, and two groups proposed and conducted optional tasks. One additional group submitted results for pooling. Cross-DB was difficult, but one group proposed a method mapping the two different semantic spaces of news articles and patents. The group proposed their approach as being applicable for various applications of IR, including mapping different user models to the IR system etc.

This was the first year of the PATENT task, and it attracted many "veterans" of NTCIR. Each group participated with its own goal for the experiments. The participants also tested unique approaches and made comparisons between different approaches, strategies and models.

Responded to the challenging task design of cross genre retrieval to the highly technical documents with extremely various length documents, Ricoh Group proposed a novel approach called “term distillation” to map different information space of newspapers and patents [11] and worked well. Follow up analysis of various retrieval models targeting to different part of patent were tested and reported elsewhere [12]

4.3 Question Answering Challenge (QAC1)

The purpose of the QAC is to encourage the development of practical QA systems in open domains and to focus on research into user interaction and information extraction. Developments of evaluation methods for question answering systems, and information resources for evaluation, are also purposes of QAC.[4].

QAC was proposed in the last NTCIR Workshop 2 [13]. We had 20 people on the organizing committee, and held four round-table meetings to discuss evaluation methods and some other problems on QA with task participants, researchers of the topic, and organizers. The QAC website is accessible at http://www.nlp.is.ritsumei.ac.jp/qac/qac1/index-j.html.

4.3.1 Subtasks

QAC1 contains three subtasks:

Task 1 System extracts five possible answers from the documents in some order. 100 questions. Document ID is required as support.

Task 2: System extracts only one set and the answers from the documents. The same 100 questions. Document ID is required as support.

Task 3: Evaluation of a series of questions. The related questions are given for 40 of the questions of Task 2.

4.3.2 Test collection

(1) Documents: The same Japanese document collections used in CLIR, Mainichi Newspapers, 1998–99.

(2) Questions: 200 for tasks 1 and 2, 40 for task 3. Approximately 900 questions were available for additional runs to enhance the test collection and to be usable for research. 60 were used for the dry run.

Answers to be extracted are the *exact answers* consisting of nouns or noun phrases. Questions are 5W1H-type questions, which are basically questions of fact—when, who, where, etc—with a variety of contents. Some of the questions have more than one answer or no answer.

4.3.3 Evaluation

Task 1: Check whether answers are correct or not and calculate the mean of reciprocal number (MRR), or the inverse number of the highest rank of the correct answers and the sum of the reciprocal numbers.

Task 2: Check whether answers are correct or not, and calculate the *F-measure* for each question and the sum of the F-measures.

Task 3: Check whether answers to related queries are correct or not, and calculate the F-measure for the branch-questions and the sum of the F-measures.

In task 1, submitted wrong answers were not given a penalty, and topics with no answer were ignored. For tasks 2 and 3 both wrong answers submitted and un-submitted right answers were penalized. In addition, in task 2 and 3, returning NUL ("no answers") gains a score. The answer can be extracted from newspaper collections as well as other knowledge sources if the answer can be found in support articles in the given document collections.

Sixteen systems from 14 research groups submitted results, and two organizers submitted the results. A scoring tool was created by the organizers. The stability of the MRR was tested, the difficulty of the questions was analyzed, and a healthy balance was indicated.

4.4 Text Summarization Challenge (TSC2)

TSC started in order for researchers in the field to collect and share text data for summarization, and to make clear the issues of evaluation measures for summarization of Japanese texts. This is the second of the series in the NTCIR and contains the new subtask of multiple document summarization [5].

4.4.1 Subtasks

TSC2 contains two subtasks:

Task A (single-document summarization): Given the texts to be summarized and summary lengths, the participants submit summaries for each text in plain text format.

Task B (multi-document summarization): Given a set of texts, the participants produce summaries of the set in plain text format. The information that was used to produce the document set, such as queries as well as summary lengths, is given to the participants.

4.4.2 Test collection

We use the same Japanese document collection as CLIR and QAC, *i.e.*, newspaper articles from the Mainichi newspaper database of 1998–1999. As key data (human-prepared summaries), we prepared the following types of summaries.

(1) *Extract-type summaries:* Experienced captioners select important sentences from each article. The summarization rates are 10%, 30%, and 50%.

(2) *Abstract-type summaries:*

(3) Summaries from more than one article: Given a set of newspaper articles selected based on a certain topic, the captioners produce free summaries for the set.

Types (1) and (2) were used as summaries from a single document, and type (3) can be used as summaries for task B. We use 30 articles for task A and 30 sets of texts (30 topics) for task B for the formal run evaluation. For each set of summaries, the summarization rates were 20% and 40%. Each document or set of documents was summarized by three different analysts.

4.4.3 Evaluation

TSC2 used subjective, intrinsic evaluation, (1) evaluation by "ranking" by human assessors who are shown several summaries and assign a rank considering content and readability, and (2) evaluation by revision—how many edits (deletions, insertions and replacements) were required by human assessors on the system-produced summary.

Nine groups conducted the task and submitted results.

4.5 Web Retrieval Task (WEB) [6]

The purpose of the WEB task is to research the retrieval of Web documents that have a structure with tags and links, and are written in Japanese or English. Web documents and the search on them contain various characteristics other than those in traditional text retrieval.

4.5.1 Subtasks

The WEB task consisted of these subtasks:

A: Retrieval for survey
  A1: Topic Retrieval
  A2: Similarity Retrieval
B: Target Retrieval
C: Optional Tasks
  C1: Search Results Classification
  C2: Speech-Driven Retrieval

Survey retrieval is a survey that aims to retrieve as many relevant documents as possible. *Target retrieval* is a search aiming for a few highly relevant documents to obtain a quick answer for the search request represented as a topic. A1 is an ordinary *ad hoc* search by topic terms, and A2 is a search by relevant documents provided as <RDOC>, a topic. Because of the unique characteristics of the document collection, we again set a free-style optional task.

Topic fields used for mandatory runs may vary according to the subtask.

4.5.2 Test collection

The document collections used were NW100G-01 (100-gigabyte) and NW10G-01 (10-gigabyte), mainly gathered from the '.jp' domain with links. The participants are only provided access to the documents at the 'Open Laboratory' in the NII (see Fig. 2 in [6]).

The topic format was specially designed to reflect the focused characteristic aspect of Web retrieval, and contains several extra fields specific to Web retrieval. Topics such as <RDOC>, known relevant documents, <USER>, information on the topic author, and <TITLE> are considered typical search queries to a search engine. These were specially formatted and used as the mandatory run. A sample topic is shown in Fig. 3 in [6].

In the relevance judgments, one-hop linked documents were also considered because we used a one-click distance model in the WEB. Judgments were made in four grades, plus selection of the "top relevant" documents. The assessors also made additional assessments of coherence or reliability.

"trec-eval", Discounted Cumulative Gain, and Weighted Mean Reciprocal Rank (MWRR), which was a newly proposed extension for multigrade judgments, were used.

Sixteen groups enrolled and seven groups submitted retrieval results. All active participants have their own unique approach and participated with their own purpose of experiment. For example, one group was a collaborative group with two different backgrounds, context-based text retrieval and link-based data mining and data warehousing. Such collaboration resulted in interesting effects on retrieval strategies.

This task was also a considerable challenge for both participants and organizers in various ways, and both participants and organizers faced many new challenges.

5 Summary

The Third NTCIR Workshop tasks include new challenges in each area of research. They can be listed as follows, although some of them were cancelled.

(1) Multilingual CLIR (CLIR).
(2) Search by Document (PATENT, WEB).
(3) Submiting passages to support the answers or relevant information in the retrieved documents (PATENT, QAC, WEB)
(4) Optional Task (PATENT, WEB).
(5) Multi-grade Relevance Judgments (CLIR, PATENT, WEB).
(6) Various types of assessment and one-click model of relevance assessment (WEB).
(7) Precision-Oriented Evaluation (QAC, WEB).
(8) New document types (WEB, PATENT).

For (1), it is our first trial of the "*CLEF mode"l* in Asia. For (4), PATENT and WEB tasks invited any research proposal from anyone interested in research using the document collections of PATENT and WEB because of the unique characteristics of these document collections.

This was the first workshop with such challenging tasks, and successive workshops have usually found a wider variety of effective strategies for tackling problems than the previous NTCIRs. In addition, "passage retrieval" providing support information as passage supporting answer or relevant information were proposed in PATENT, QAC and WEB, however all of these were cancelled. However, that direction is one of the natural extensions of research in information access technologies. These are among the proposed extensions for future workshops.

The results of MLIR in CLIR showed there is still considerable scope for investigation of East Asian language and cross-language approaches with English and European languages. We also continue to the direction as "information access" as well.

References

[1] NTCIR Project: http://research.nii.ac.jp/ntcir/.

[2] Chen, K.H., Chen, H.H., Kishida, K., Kuriyama, K., Kando, N., Lee, S., Myaeng, S.H., Eguchi, K., Kim, H. "Overview of CLIR Task at the Third NTCIR Workshop". *In NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo, Japan, September 2001 – October 2002,* ISBN 4-86049-016-9. (This volume) [Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-OV-CLIR-ChenK.rev.pdf].

[3] Iwayama, M., Fujii, A., Kando, N., Takano, A. "Overview of Patent Retrieval Task at NTCIR-3". *In NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo, Japan, September 2001 – October 2002,* ISBN 4-86049-016-9. (This volume) [Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-OV-PATENT-IwayamaM.pdf]

[4] Fukumoto, J., Kato, T., Masui, F. "Question Answering Challenge (QAC-1): Question answering evaluation at NTCIR Workshop 3", *In NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo, Japan, September 2001 – October 2002,* ISBN 4-86049-016-9. (This volume) [Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-OV-QAC-FukumotoJ.pdf]

[5] Fukusima, T., Okumura, M., Nanba, H. "Text summarization Challenge 2: Text summarization evaluation at NTCIR Workshop3". *In NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo, Japan, September 2001 – October 2002,* ISBN 4-86049-016-9. (This volume) [Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html]

[6] Eguchi, K., Oyama, K., Ishida, E., Kando, N., Kuriyama, K. "Overview of Web Retrieval Task at the Third NTCIR Workshop", *In NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo, Japan, September 2001 – October 2002,* ISBN 4-86049-016-9. (This volume) [Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-OV-WEB-EguchiK.pdf]

[7] TREC: Text REtrieval Conference: URL: http://trec.nist.gov/

[8] *NTCIR Workshop 1: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, *Tokyo, Japan, 30 Aug.–1 Sep., 1999.* 487 p. ISBN4-924600-77-6. [Available : http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings/index.html]

[9] Kageura, K., Koyama, T. Guest Editors. Special issue of *Terminology*, Vol.6, No.2, 2002.

[10] *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, *Tokyo, Japan, June 2000–March 2001.* 380 p. ISBN4-924600-96-2. [Available : http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/index.html].

[11] Itoh, H., Mano, H., Ogawa, Y. "Term Distillation for Cross DB Retrieval", *In NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo, Japan, September 2001 – October 2002,* ISBN 4-86049-016-9. (This volume) [Available: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-PATENT-ItohH.pdf].

[12] Iwayama, M., Fujii, A., Kando, N., Marukawa, Y. "An empirical study on retrieval models for different document genres: Patents and newspaper articles", I*n Proceedings of the 26th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), July 28-Aug., 2003, Tronto, Canada*, (to appear)

[13] Fukumoto, J., Kato, T. "An Overview of Question and Answering Challenge (QAC) in the next NTCIR Workshop" In *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, *Tokyo, Japan, June 2000–March 2001.* pp.375-377 ISBN4-924600-96-2. [Available : http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/fukumoto.pdf]

[14] Cleverdon, CW, "The Cranfield tests on index language devices", *Aslib Proceedings*, 19, p.173-192, 1967.

[15] Salton, G. ed. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, 1971.

[16] Sparck Jones, K., Rijsbergen, C.J. *Report on the need for and provision of an 'ideal' information retrieval test collection*, Computer laboratory, University of Cambridge, 1975 (BLRDD Report).

[17] Harman, D. "The Development and Evolution of TREC and DUC", *In NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo, Japan, September 2001 – October 2002,* ISBN 4-86049-016-9. (This volume) [Available:

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-INV-HarmanD.pdf]

[18] CLEF: Cross-Language Evaluation Forum [http://www.iei.pi.cnr.it/DELOS/CLEF].

[19] Saracevic, T. "Relevance reconsidered '96", In *Proceedings of the 2nd Conference on Library and Information Science (CoLIS-2*), Copenhagen, Denmark, Oct. 1996, p.201–218.

[20] Mizzaro, S. "Relevance: The whole history" *Journal of the American Society for Information Science,* Vol.48, No.9, pp.810–832, 1996.

[21] Schamber, L. "Relevance and information behavior", *In Annual Review of Information Science and Technology,* Vol.29, pp.29-48, 1994.

[22] Nozue, T., Kando, N. "Primary considerations in the concept of relevance: Relevance judgement of NTCIR". *IPSJ SIG Notes, 99-FI-53,* Vol.99, No.20, March 1999, p. 49–56. (in Japanese with English abstract).

[23] Kando, N. "Relevance Re-Examined: In the Context of Information Retrieval System Testing", Presented at *International Symposium on the Logic of Real-World Interaction (LRWI 2002)*, Jan. 30–31, 2002.

[24] Jarvelin, K., Kekalainen, J. IR evaluation methods for retrieving highly relevant documents. *In Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Athns, 2000,* pp.41-48.

[25] Spink, A., Greisdorf, H. "Regions and levels:". *Journal of the American Society for Information Sciences*, Vol.52, No.2, pp.161–173, 2001.

[26] Kando, N., Kuriyama, K., Yoshioka, M. "Evaluation based on multi-grade relevance judgments". *IPSJ SIG Notes*, Vol.2001-FI-63, pp.105–112, July 2001. (in Japanese with English abstract).

[27] Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C-Y., Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski1, T., Voorhees, E., Weishedel, R. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). *NIST DUC Vision and Roadmap Documents* [Available: http://www-nlpir.nist.gov/projects/duc/roadmap.html, 2001].

[28] Voorhees, E.M. "Overview of the TREC-9 Question Answering Track", *In Proceedings of the Ninth Text Retrieval Conference (TREC-9), 2001*

[29] Voorhees, E.M., Tice, D.M. "Building a Question Answering Test Collection",, *In Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Athens, 2000*, pp.200–207.

[30] Yoshioka, M., Kuriyama, K., Kando, N.: "Analysis on the usage of Japanese segmented texts in the NTCIR Workshop 2." In *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, *Tokyo, Japan, June 2000–March 2001.* pp.291-296 ISBN4-924600-96-2. [Available : http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/yoshioka.pdf]

[31] Kando, N, Nozue, T., Kuriyama, K., Oyama, K., "NTCIR-1: Its policy and practice", *IPSJ SIG Notes*, Vol.99, No.20, pp.33–40, 1999. (in Japanese with English abstract).

[32] Kuriyama, K., Nozue, T., Kando, N., Oyama, K.: "Pooling for a large scale test collection: Analysis of the search results for the pre-test of the NTCIR-1 Workshop", *IPSJ SIG Notes*, Vol.99-FI-54, pp.25–32 May, 1999 (in Japanese with English abstract).

[33] Kuriyama, K., Kando, K. "Construction of a large scale test collection: Analysis of the training topics of the NTCIR-1", *IPSJ SIG Notes*, Vol.99-FI-55, pp.41—48, July 1999. (in Japanese with English abstract).

[34] Kando, N., Eguchi, K., Kuriyama, K., "Construction of a large scale test collection: Analysis of the test topics of the NTCIR-1", In *Proceedings of IPSJ Annual Meeting* (in Japanese). pp.3-107 – 3-108, 30 Sep.–3 Oct. 1999.

[35] Kuriyama, K., Yoshioka, M., Kando, N., "Effect of cross-lingual pooling". In *NTCIR Workshop 2: Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, June 2000–March 2001. pp.297-310. （ISBN：4-924600-96-2. [Available : http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/kuriyama.pdf]

[36] Kuriyama, K., Kando, N. "Pooling for a large-scale test collection : An analysis of the search results from the first NTCIR Workshop." *Information Retrieval*, Vol.5, No.1, pp.41-59, January 2002

[37] Eguchi, K., Kuriyama, K., Kando, N. "Analysis of the topic difficulty for NTCIR (NACSIS Test Collection for information Retrieval Systems)". In *Proceedings of the 3rd International Conference of Asian Digital Libraries, Seoul, Korea, December, 2000,* p.231-238

[38] Voorhees, E.M., "Variations in relevance judgments and the measurement of retrieval effectiveness", In *Proceedings of SIGIR '98*, pp.315–323.

[39] Buckley, C. trec-eval IR evaluation package. Available from ftp://ftp.cs.cornell.edu/pub/smart/.

[40] Kando, N. "Towards real multilingual information discovery and access". Presented at *ACM Digital Libraries and ACM-SIGIR Joint Workshop on Multilingual Information Discovery and Access (MIDAS). Berkeley, CA, 15 August, 1999.*